UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/796,644	03/09/2004	Scott Meredith	M61.12-0602	2206

27366 7590 08/22/2006

WESTMAN CHAMPLIN (MICROSOFT CORPORATION)
SUITE 1400
900 SECOND AVENUE SOUTH
MINNEAPOLIS, MN 55402-3319

EXAMINER	
LOVEL, KIMBERLY M	

ART UNIT	PAPER NUMBER
2167	

DATE MAILED: 08/22/2006

Please find below and/or attached an Office communication concerning this application or proceeding.

Office Action Summary	Application No.	Applicant(s)	
	10/796,644	MEREDITH ET AL.	
	Examiner	Art Unit	
	Kimberly Lovel	2167	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *09 March 2004*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-25* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-25* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *09 March 2004* is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☐ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08) Paper No(s)/Mail Date *9/26/05 1/9/06*.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____.
5) ☐ Notice of Informal Patent Application (PTO-152)
6) ☐ Other: _____.

DETAILED ACTION

1. Claims 1-25 are rejected.

Information Disclosure Statement

2. The information disclosure statements (IDS) submitted on 26 September 2005 and 9 January 2006 were filed after the mailing date of the application on 9 March 2004. The submission is in compliance with the provisions of 37 CFR 1.97. Accordingly, the information disclosure statements are being considered by the examiner.

Claim Rejections - 35 USC § 102

3. The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

> A person shall be entitled to a patent unless –
>
> (b) the invention was patented or described in a printed publication in this or a foreign country or in public use or on sale in this country, more than one year prior to the date of application for patent in the United States.

4. Claims 1 and 6 are rejected under 35 U.S.C. 102(b) as being anticipated by the article "IntelliClean: A Knowledge-Based Intelligent Data Cleaner" by Lee et al (hereafter Lee et al).

Referring to claim 1, Lee et al disclose a method of compressing a log of linguistic data [Patient dataset] (see page 293, right column, section 5.2: Cleaning the patient dataset), the log having a plurality of linguistic strings [records] (see page 293, right column, section 5.2, first paragraph, line 1), each string being including at least two

tokens [60 fields] (see page 293, right column, section 5.2, first paragraph, line 2), the method comprising:

applying a compression operation [pre-processing operation – Pre-processing cleans dirty data. According to lines 4-6 of the Introduction of Lee et al, dirty data includes misuse of abbreviations, data entry mistakes, control information, missing fields, spelling, outdated codes, etc. These examples are similar to those given by page 11, line 23 – page 14, line 12 of the applicants' specification wherein the different levels of compression are defined. Therefore, pre-processing is considered to be analogous to compression] to each string (page 291, right column, first paragraph);

determining if any two strings match each other after the compression operation (Figure 2 and page 291, right column, second paragraph, lines 1-6 – Duplicate Identification Rules); and

removing one of the two matching strings from the log (Figure 2 and page 291, right column, second paragraph, lines 1-3 and 7-12 – Merge/Purge Rules; only the record with the least number of empty fields is to be kept in a group of duplicate records).

Referring to claim 6, Lee et al disclose the method of claim 1, wherein the compression operation [pre-processing operation] is token-based (see page 290, right column, section 3: Related Works, lines 1-5 – pre-processing performed at the token level).

Claim Rejections - 35 USC § 103

5. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

This application currently names joint inventors. In considering patentability of the claims under 35 U.S.C. 103(a), the examiner presumes that the subject matter of the various claims was commonly owned at the time any inventions covered therein were made absent any evidence to the contrary. Applicant is advised of the obligation under 37 CFR 1.56 to point out the inventor and invention dates of each claim that was not commonly owned at the time a later invention was made in order for the examiner to consider the applicability of 35 U.S.C. 103(c) and potential 35 U.S.C. 102(e), (f) or (g) prior art under 35 U.S.C. 103(a).

6. Claims 2-4 are rejected under 35 U.S.C. 103(a) as being unpatentable over the article "IntelliClean: A Knowledge-Based Intelligent Data Cleaner" by Lee et al as applied to claim 1 above, and further in view of the article "Mining Generalized Patterns from Web Logs" by Ling et al (hereafter Ling et al).

Referring to claim 2, Lee et al disclose a log. However, Lee et al fail to explicitly disclose the further limitation wherein the log is a log of queries. Ling et al disclose logs used in data mining (see abstract), including the further limitation wherein the log is a log of queries [logs that keep track of user queries] (see abstract, lines 1-2) since data

cleaning is an important for data mining and it yields a good data set for knowledge retrieval.

One of ordinary skill in the art would have been motivated to replace the log mentioned by Lee et al with the log of queries mentioned by Ling et al. One would have been motivated to do so since data cleaning is important for data mining and it yields a good data set for knowledge retrieval (Lee et al: see page 290, right column, section 3: Related Works).

Referring to claim 3, the combination of Lee et al and Ling et al (hereafter Lee/Ling) discloses the method of claim 2, wherein the queries are queries relative to a help function (Ling et al: see abstract, lines 12-14 – the queries aid the web editors in discovering topics in which the users are interested, which is considered to represent the help function).

Referring to claim 4, Lee/Ling discloses the method of claim 3, wherein the help-related queries are relative to a computer system [search engine] (Ling et al: see abstract, lines 9-12).

7. Claims 14-16 and 18 are rejected under 35 U.S.C. 103(a) as being unpatentable over the article "IntelliClean: A Knowledge-Based Intelligent Data Cleaner" by Lee et al in view of the article "Mining Generalized Patterns from Web Logs" by Ling et al.

Referring to claim 14, Lee et al disclose a system for compressing a log [Patient dataset] (see page 293, right column, section 5.2: Cleaning the Patient dataset) having a plurality of linguistic strings [records] (see page 293, right column, section 5.2, first

paragraph, line 1), each string having a plurality of tokens [60 fields] (see page 293, right column, section 5.2, first paragraph, line 2), the system comprising:

an input for receiving a log [pre-processing stage] (see Figure 2);

memory for storing the log [Original Records repository] (see Figure 2);

a processor [expert system engine] for applying at least one compression operation [pre-processing operation] to each string (page 291, right column, first paragraph), and for scanning the modified strings to determine if any match each other (Figure 2 and page 291, right column, second paragraph, lines 1-6 – Duplicate Identification Rules) so that one of the matching strings can be removed (Figure 2 and page 291, right column, second paragraph, lines 1-3 and 7-12 – Merge/Purge Rules; only the record with the least number of empty fields is to be kept in a group of duplicate records); and

an output [the cleansed data records repository] for providing a compressed log once the removal is complete (see Figure 2).

However, Lee et al fail to explicitly disclose the further limitation wherein the log is query log. Ling et al disclose logs used in data mining (see abstract), including the further limitation wherein the log is a query log [logs that keep track of user queries] (see abstract, lines 1-2) since data cleaning is an important for data mining and it yields a good data set for knowledge retrieval.

One of ordinary skill in the art would have been motivated to replace the log mentioned by Lee et al with the log of queries mentioned by Ling et al. One would have been motivated to do so since data cleaning is important for data mining and it yields a

good data set for knowledge retrieval (Lee et al: see page 290, right column, section 3: Related Works).

Referring to claim 15, hereafter Lee/Ling discloses the system of claim 14, wherein the queries are queries relative to a help function (Ling et al: see abstract, lines 12-14 – the queries aid the web editors in discovering topics in which the users are interested, which is considered to represent the help function).

Referring to claim 16, Lee/Ling discloses the system of claim 15, wherein the help-related queries are relative to a computer system [search engine] (Ling et al: see abstract, lines 9-12).

Referring to claim 18, Lee et al disclose the system of claim 14, wherein the compression operation [pre-processing operation] is token-based (see page 290, right column, section 3: Related Works, lines 1-5 – pre-processing performed at the token level).

8. Claims 5 and 17 are rejected under 35 U.S.C. 103(a) as being unpatentable over the article "IntelliClean: A Knowledge-Based Intelligent Data Cleaner" by Lee et al as applied respectively to claims 1 and 14 above, and further in view of the article "Better Rules, Fewer Features: A Semantic Approach to Selecting Features from Text" by Blake et al (hereafter Blake et al).

Referring to claim 5, Lee et al disclose a compression operation. However, Lee et al fail to explicitly disclose the further limitation wherein the compression operation is character-based. Blake et al disclose compression based on a comparison at three

different levels (see abstract), including the further limitation wherein the compression operation is character-based [stopwords] (see page 62, right-hand column, lines 5-19 – removing the stopwords is considered to represent an example of character-based compression) in order to increase the accuracy of determining if two strings are duplicates.

One of ordinary skill in the art would have been motivated to use the feature of Blake et al for compressing strings that are character-based with Lee et al's compression operation. One would have been motivated to do so in order to increase the accuracy of determining if two strings are duplicates.

Referring to claim 17, Lee et al disclose a compression operation. However, Lee et al fail to explicitly disclose the further limitation wherein the compression operation is character-based. Blake et al disclose compression based on a comparison at three different levels (see abstract), including the further limitation wherein the compression operation is character-based [stopwords] (see page 62, right-hand column, lines 5-19 – removing the stopwords is considered to represent an example of character-based compression) in order to increase the accuracy of determining if two strings are duplicates.

One of ordinary skill in the art would have been motivated to use the feature of Blake et al for compressing strings that are character-based with Lee et al's compression operation. One would have been motivated to do so in order to increase the accuracy of determining if two strings are duplicates.

9. Claims 7-8 and 19-20 are rejected under 35 U.S.C. 103(a) as being unpatentable over the article "IntelliClean: A Knowledge-Based Intelligent Data Cleaner" by Lee et al as applied respectively to claims 1 and 14 above, and further in view of the article "Faster Algorithm of String Comparison" by Yang et al (hereafter Yang et al).

Referring to claim 7, Lee et al disclose a compression operation. However, Lee et al fail to disclose the further limitation wherein the compression operation is subsumption. Yang et al disclose improving the calculation of similarity (see right column, last paragraph, lines 1-4), including the further limitation wherein the compression operation is subsumption (see page 123, right column, section: Proposed New Field Similarity) in order to increase the accuracy of determining if two strings are duplicates.

One of ordinary skill in the art would have been motivated to use the feature of Yang et al for using subsumption to compress strings with Lee et al's compression operation. One would have been motivated to do in order to increase the accuracy of determining if two strings are duplicates.

Referring to claim 8, the combination of Lee et al and Yang et al discloses the method of claim 7, wherein subsumption includes applying an impossibility condition to selectively compute edit distance [SIM formula] (see page 123, right column, section: Proposed New Field Similarity).

Referring to claim 19, Lee et al disclose a compression operation. However, Lee et al fail to disclose the further limitation wherein the compression operation is subsumption. Yang et al disclose improving the calculation of similarity (see right

column, last paragraph, lines 1-4), including the further limitation wherein the compression operation is subsumption (see page 123, right column, section: Proposed New Field Similarity) in order to increase the accuracy of determining if two strings are duplicates.

One of ordinary skill in the art would have been motivated to use the feature of Yang et al for using subsumption to compress strings with Lee et al's compression operation. One would have been motivated to do in order to increase the accuracy of determining if two strings are duplicates.

Referring to claim 20, the combination of Lee et al and Yang et al discloses the system of claim 19, wherein subsumption includes applying an impossibility condition to selectively compute edit distance [SIM formula] (see page 123, right column, section: Proposed New Field Similarity).

10. Claims 9-13 and 21-25 are rejected under 35 U.S.C. 103(a) as being unpatentable over the article "IntelliClean: A Knowledge-Based Intelligent Data Cleaner" by Lee et al as applied respectively to claims 1 and 14 above, and further in view of the article "From Data Mining to Knowledge Discovery in Databases" by Fayyad et al (hereafter Fayyad et al).

Referring to claim 9, Lee et al discloses a method of compressing a log of linguistic data, the log having a plurality of linguistic strings, each string being including at least two tokens. However, Lee et al fail to explicitly disclose the further limitations of applying a second compression operation to each string; determining if any two strings

match each other after the second compression operation; and removing one of the two matching strings from the log. Fayyad discloses the method of data mining and knowledge discovery (see abstract) including the further limitations of applying a second compression operation to each string; determining if any two strings match each other after the second compression operation; and removing one of the two matching strings from the log (see page 42, left column, lines 1-33) in order to provide a further compression of the strings.

One of ordinary skill in the art would have been motivated to use the feature of Yang et al for applying a second compression operation with Lee et al's first compression operation. One would have been motivated to do in order in order to provide a further compression of the strings.

Referring to claim 10, the combination of Lee et al and Fayyad et al (hereafter Lee/Fayyad) discloses the method of claim 9, wherein the first compression operation is character-based and the second compression operation is token based (Fayyad et al: see page 42, left column, lines 1-33).

Referring to claim 11, Lee/Fayyad discloses the method of claim 10, and further comprising applying subsumption after the second compression operation is complete (Fayyad et al: see page 42, left column, lines 1-33).

Referring to claim 12, Lee/Fayyad discloses the method of claim 11, wherein the subsumption operation is repeated for the log (Fayyad et al: see page 42, left column, lines 1-33).

Referring to claim 13, Lee/Fayyad discloses the method of claim 1, and further comprising training a statistical process with the compressed log (Fayyad et al: see page 42, left column, lines 1-33).

Referring to claim 21, Lee et al discloses a system of compressing a log of linguistic data, the log having a plurality of linguistic strings, each string being including at least two tokens. However, Lee et al fail to explicitly disclose the further limitations of applying a second compression operation to each string; determining if any two strings match each other after the second compression operation; and removing one of the two matching strings from the log. Fayyad discloses a system of data mining and knowledge discovery (see abstract) including the further limitations of applying a second compression operation to each string; determining if any two strings match each other after the second compression operation; and removing one of the two matching strings from the log (see page 42, left column, lines 1-33) in order to provide a further compression of the strings.

One of ordinary skill in the art would have been motivated to use the feature of Yang et al for applying a second compression operation with Lee et al's first compression operation. One would have been motivated to do in order in order to provide a further compression of the strings.

Referring to claim 22, the combination of Lee et al and Fayyad et al (hereafter Lee/Fayyad) discloses the system of claim 21, wherein the first compression operation is character-based and the second compression operation is token based (Fayyad et al: see page 42, left column, lines 1-33).

Referring to claim 23, Lee/Fayyad discloses the system of claim 22, and further comprising applying subsumption after the second compression operation is complete (Fayyad et al: see page 42, left column, lines 1-33).

Referring to claim 24, Lee/Fayyad discloses the system of claim 23, wherein the subsumption operation is repeated for the log (Fayyad et al: see page 42, left column, lines 1-33).

Referring to claim 25, Lee/Fayyad discloses the system of claim 14, and further comprising training a statistical process with the compressed log (Fayyad et al: see page 42, left column, lines 1-33).

Contact Information

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Kimberly Lovel whose telephone number is (571) 272-2750. The examiner can normally be reached on 8:00 - 4:00.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, John Cottingham can be reached on (571) 272-7079. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

Kimberly Lovel
Examiner
Art Unit 2167

18 Aug 2006
kml

JOHN COTTINGHAM
SUPERVISORY PATENT EXAMINER
TECHNOLOGY CENTER 2100

18 August 2006